What is claimed is:

1. A computer implemented method for determining the direction of a consensus sequence of a cluster of sequences with contradictions in directions comprising:

determining the probability (*b*) that the contradictions are explained by random errors according to a statistical model and the weighted number of contradictory sequences in the cluster; and

defining the direction of majority of the sequences as the direction of the consensus sequence if the probability is the same as or greater than a threshold value (*T*) and x≠n/2.

2. The method of Claim 1 wherein the statistical model is a binomial distribution and the probability is calculated as follows:

$$b(x;n,p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

wherein *n* is the weighted number of the sequences in the cluster; *p* is the probability of random errors resulting in the contradictions; and *x* is the number of the contradictory sequences.

3. The method of Claim 2 wherein CDS and mRNA sequences carry a higher weight than 5' EST or 3' EST; directionless EST carrys a weight of 0.

4. The method of Claim 2 wherein the weights to different types of sequences are the same.

5. The method of Claim 2 wherein the threshold value is around 0.001.

6. The method of Claim 2 wherein the threshold value is around 0.002.

7. The method of Claim 2 wherein the threshold value is around 0.003.

8. The method of Claim 2 further comprising defining the direction of majority of the sequences as the direction of the consensus sequence if the probability is lower than the threshold value and $x \leq n^*(P_t)$.

9. The method of Claim 8 further comprising further subclustering for the minority direction and majority direction if the probability is smaller than the threshold value and $x>n^*P$.

10. The method of Claim 9 wherein the p is between 0.03-0.10.

11. The method of Claim 10 wherein the p is around 0.06.

12. The method of Claim 11 wherein the p is determined according to binomial frequency distribution of contradictory sequences in a plurality of clusters or subclusters of sequences.

13. A method of selecting sequences for designing a probe array comprising:
cleaning raw sequences;
refining clusters of the raw sequences; and
generating candidate design sequences, wherein the candidate design sequences are exemplar or consensus sequences of the clusters.

14. The method of Claim 13 wherein the cleaning comprises
removing withdrawn sequences;
screening and filtering and masking raw sequences; and
triming terminal ambiguous sequence regions.

15. The method of Claim 13 wherein the refining includes two level clustering.

16. The method of Claim 13 wherein the generating comprises:
selecting exemplary sequences.

17. The method of Claim 16 wherein the generating comprises:
generating alignments of sequences within clusters;
calling consensus sequence bases according to consensus calling rules; and
determining consensus sequence direction.

18. The method of Claim 17 wherein the determining comprises defining the direction of sequences in the clusters as the consensus sequence direction if there is no contradictory sequence directions.

19. The method of Claim 18 wherein the determining further comprises
determining the probability (*b*) that the contradictions are explained by random errors according to a statistical model and the weighted number of contradictory sequences in the cluster; and
defining the direction of majority of the sequences as the direction of the consensus sequence if the probability is the same as or greater than a threshold value (*T*) and x≠n/2.

20. The method of Claim 19 wherein the statistical model is a binomial distribution and the probability is calculated as follows:

$$b(x;n,p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

wherein *n* is the weighted number of the sequences in the cluster; *p* is the probability of random errors resulting in the contradictions; and *x* is the number of the contradictory sequences.

21. The method of Claim 20 wherein CDS and mRNA sequences carry a higher weight than 5' EST or 3' EST; directionless EST carrys a weight of 0.

22. The method of Claim 21 wherein the weights to different types of sequences are the same.

23. The method of Claim 22 wherein the threshold value is around 0.001.

24. The method of Claim 22 wherein the threshold value is around 0.002.

25. The method of Claim 22 wherein the threshold value is around 0.003.

26. The method of Claim 22 further comprising defining the direction of majority of the sequences as the direction of the consensus sequence if the probability is lower than the threshold value and $x \leq n^*(P_t)$.

27. The method of Claim 26 further comprising further subclustering for the minority direction and majority direction if the probability is smaller than the threshold value and $x>n^*P$.

28. The method of Claim 27 wherein the p is between 0.03-0.10.

29. The method of Claim 27 wherein the p is around 0.06.

30. The method of Claim 27 wherein the p is determined according to binomial frequency distribution of contradictory sequences in a plurality of clusters or subclusters of sequences.